

# Walking on a Graph with a Magnifying Glass

## Stratified Sampling via Weighted Random Walks

Maciej Kurant, Minas Gjoka, Carter T. Butts, Athina Markopoulou  
 University of California, Irvine  
 {mkurant, mgjoka, buttsc, athina}@uci.edu

### ABSTRACT

Our objective is to sample the node set of a large unknown graph via crawling, to accurately estimate a given metric of interest. We design a random walk on an appropriately defined weighted graph that achieves high efficiency by preferentially crawling those nodes and edges that convey greater information regarding the target metric. Our approach begins by employing the theory of stratification to find optimal node weights, for a given estimation problem, under an independence sampler. While optimal under independence sampling, these weights may be impractical under graph crawling due to constraints arising from the structure of the graph. Therefore, the edge weights for our random walk should be chosen so as to lead to an equilibrium distribution that strikes a balance between approximating the optimal weights under an independence sampler and achieving fast convergence. We propose a heuristic approach (stratified weighted random walk, or S-WRW) that achieves this goal, while using only limited information about the graph structure and the node properties. We evaluate our technique in simulation, and experimentally, by collecting a sample of Facebook college users. We show that S-WRW requires 13-15 times fewer samples than the simple re-weighted random walk (RW) to achieve the same estimation accuracy for a range of metrics.

### 1. INTRODUCTION

Many types of online networks, such as online social networks (OSNs), Peer-to-Peer (P2P) networks, or the World Wide Web (WWW), are measured and studied today via sampling techniques. This is due to several reasons. First, such graphs are typically too large to measure in their entirety, and it is desirable to be able to study them based on a small but representative sample. Second, the information pertaining to these networks is often hard to obtain. For example, OSN service providers have access to all information in their user base, but rarely make this information publicly available.

There are many ways a graph can be sampled, *e.g.*, by sampling nodes, edges, paths, or other substructures [23, 27]. Depending on our measurement goal, the elements with different properties may have *different importance* and should be sampled with a different probability. For example, Fig. 1(a) depicts the world's population, with residents of China (1.3B people) represented by blue nodes, of the Vatican (800 people) by black nodes, and all other nation-

alities represented by white nodes. Assume that we want to compare the median income in China and Vatican. Taking a uniform sample of size 100 from the entire world's population is ineffective, because most of the samples will come from countries other than China and Vatican. Even restricting our sample to the union of China and Vatican will not help much, as our sample is unlikely to include any Vatican resident. In contrast, uniformly sampling 50 Chinese and 50 Vaticanese residents would be much more accurate with the same sampling budget.

This type of problem has been widely studied in the statistical and survey sampling literature. A commonly used approach is *stratified sampling* [12, 28, 34], where nodes (*e.g.*, people) are partitioned into a set of non-overlapping *categories* (or strata). The objective is then to decide how many independent draws to take from each category, so as to minimize the uncertainty of the resulting measurement. This effect can be achieved in expectation by a weighted independence sampler (WIS) with appropriately chosen sampling probabilities  $\pi^{\text{WIS}}$ . In our example, WIS samples Vatican residents with much higher probabilities than Chinese ones, and avoids completely the rest of the world, as illustrated in Fig. 1(b).

However, WIS, as every independence sampler, requires a sampling frame, *i.e.*, a list of all elements we can sample from (*e.g.*, a list of all Facebook users). This information is typically not available in today's online networks. A feasible alternative is *crawling* (also known as exploration or link-trace sampling). It is a graph sampling technique in which we can see the neighbors of already sampled users and make a decision on which users to visit next.

In this paper, we study how to perform stratified sampling through graph crawling. We illustrate the key idea and some of the challenges in Fig. 1. Fig. 1(c) depicts a social network that connects the world's population. A simple random walk (RW) visits every node with frequency proportional to its degree, which is reflected by the node size. In this particular example, for a simplicity of illustration, all nodes have the same degree equal to 3. As a result, RW is equivalent to the uniform sample of the world's population, and faces exactly the same problems of wasting resources, by sampling all nodes with the same probability.

We address these problems by appropriately setting the edge weights and then performing a random walk on the weighted graph, which we refer to as *weighted random walk* (WRW). One goal in setting the weights is to mimic the WIS-optimal sampling probabilities  $\pi^{\text{WIS}}$  shown in Fig. 1(b). However, such a WRW might perform poorly due to poten-

\* This is an extended version of a paper with the same title presented at *SIGMETRICS'11*. This work was supported by SNF grant PBELP2-130871, Switzerland, and by the NSF CDI Award 1028394, USA.

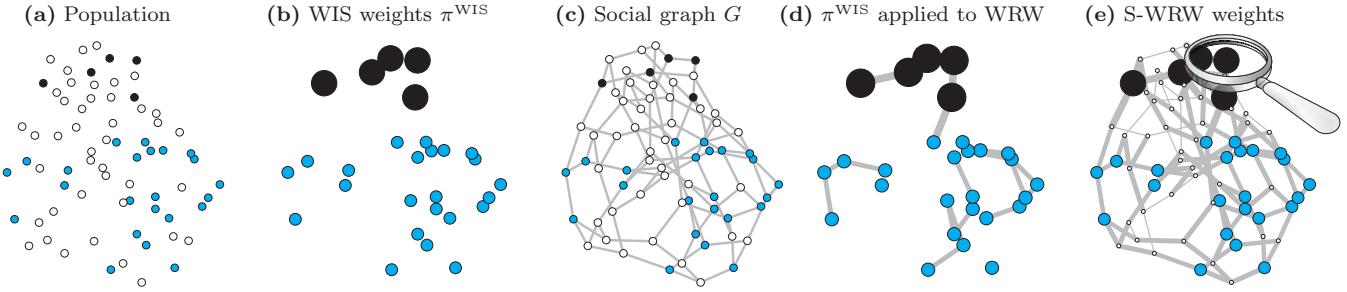


Figure 1: Illustrative example. Our goal is to compare the blue and black subpopulations (e.g., with respect to their median income) in population (a). Optimal independence sampler, WIS (b), over-samples the black nodes, under-samples the blue nodes, and completely skips the white nodes. A naive crawling approach, RW (c), samples many irrelevant white nodes. WRW that enforces WIS-optimal probabilities may result in poor or no convergence (d). S-WRW (e) strikes a balance between the optimality of WIS and fast convergence.

tially slow mixing. In our example, it will not even converge because the underlying weighted graph is disconnected, as shown in Fig. 1(d). Therefore, the edge weights under WRW (which determine the equilibrium distribution  $\pi^{\text{WRW}}$ ) should be chosen in a way that strikes a balance between the optimality of  $\pi^{\text{WIS}}$  and fast convergence.

We propose *Stratified Weighted Random Walk* (S-WRW), a practical heuristic that effectively strikes such a balance. We refer to our approach as “walking on the graph with a magnifying glass”, because S-WRW over-samples more relevant parts of the graph and under-samples less relevant ones. In our example, S-WRW results in the graph presented in Fig. 1(e). The only information required by S-WRW are the categories of neighbors of every visited node, which is typically available in crawlable online networks, such as Facebook. S-WRW uses two natural and easy-to-interpret parameters, namely: (i)  $f_{\ominus}$ , which controls the fraction of samples from irrelevant categories and (ii)  $\gamma$ , which is the maximal resolution of our magnifying glass, with respect to the largest relevant category.

The main contributions of this paper are the following.

- We propose to improve the efficiency of crawling-based graph sampling methods, by performing a stratified weighted random walk that takes into account not only the graph structure but also the node properties that are relevant to the measurement goal.
- We design and evaluate S-WRW, a practical heuristic that sets the edge weights and operates with limited information.
- As a case study, we apply S-WRW to sample Facebook and estimate the sizes of colleges. We show that S-WRW requires 13-15 times fewer samples than a simple random walk for the same estimation accuracy.

The outline of the rest of the paper is as follows. Section 2 summarizes the most popular graph sampling techniques, including sampling by exploration. Section 3 presents classical stratified sampling. Section 4 combines stratified sampling with graph exploration, presenting a unified WRW approach that takes into account both network structure and node properties; various trade-offs and practical issues are discussed and an efficient heuristic (S-WRW) is proposed based on the insights. Section 5 presents simulation results. Section 6 presents an implementation of S-WRW for the problem of estimating the college friendship graph on Face-

book. Section 7 presents related work. Section 8 concludes the paper.

## 2. SAMPLING TECHNIQUES

### 2.1 Notation

We consider an undirected, static,<sup>1</sup> graph  $G = (V, E)$ , with  $N = |V|$  nodes and  $|E|$  edges. For a node  $v \in V$ , denote by  $\deg(v)$  its degree, and by  $\mathcal{N}(v) \subset V$  the list of neighbors of  $v$ . A graph  $G$  can be weighted. We denote by  $w(u, v)$  the weight of edge  $\{u, v\} \in E$ , and by

$$w(u) = \sum_{v \in \mathcal{N}(u)} w(u, v) \quad (1)$$

the weight of node  $u \in V$ . For any set of nodes  $A \subseteq V$ , we define its volume  $\text{vol}(A)$  and weight  $w(A)$ , respectively, as

$$\text{vol}(A) = \sum_{v \in A} \deg(v) \quad \text{and} \quad w(A) = \sum_{v \in A} w(v). \quad (2)$$

We will often use

$$f_A = \frac{|A|}{|V|} \quad \text{and} \quad f_A^{\text{vol}} = \frac{\text{vol}(A)}{\text{vol}(V)} \quad (3)$$

to denote the relative size of  $A$  in terms of the number of nodes and the volumes, respectively.

**Sampling.** We collect a sample  $S \subseteq V$  of  $n = |S|$  nodes.  $S$  may contain multiple copies of the same node, *i.e.*, the sampling is with replacement. In this section, we briefly review the techniques for sampling nodes from graph  $G$ . We also present the weighted random walk (WRW) which is the basic building block for our approach.

### 2.2 Independence Sampling

**Uniform Independence Sampling (UIS)** samples the nodes directly from the set  $V$ , with replacements, uniformly and independently at random, *i.e.*, with probability

$$\pi^{\text{UIS}}(v) = \frac{1}{N} \quad \text{for every } v \in V. \quad (4)$$

**Weighted Independence Sampling (WIS)** is a weighted version of UIS. WIS samples the nodes directly from the

<sup>1</sup>Sampling dynamic graphs is currently an active research area [35,40,42], but out of the scope of this paper.

set  $V$ , with replacements, independently at random, but with probabilities proportional to node weights  $w(v)$ :

$$\pi^{\text{WIS}}(v) = \frac{w(v)}{\sum_{u \in V} w(u)}. \quad (5)$$

In general, UIS and WIS are not possible in online networks because of the lack of sampling frame. For example, the list of all user IDs may not be publicly available, or the user ID space may be too sparsely allocated. Nevertheless, we present them as baseline for comparison with the random walks.

### 2.3 Sampling via Crawling

In contrast to independence sampling, the crawling techniques are possible in many online networks, and are therefore the main focus of this paper.

**Simple Random Walk (RW)** [29] selects the next-hop node  $v$  uniformly at random among the neighbors of the current node  $u$ . In a connected and aperiodic graph, the probability of being at the particular node  $v$  converges to the stationary distribution

$$\pi^{\text{RW}}(v) = \frac{\deg(v)}{2 \cdot |E|}. \quad (6)$$

**Metropolis-Hastings Random Walk (MHRW)** is an application of the Metropolis-Hastings algorithm [30] that modifies the transition probabilities to converge to a desired stationary distribution. For example, we can achieve the uniform stationary distribution

$$\pi^{\text{MHRW}}(v) = \frac{1}{N} \quad (7)$$

by randomly selecting a neighbor  $v$  of the current node  $u$  and moving there with probability  $\min(1, \frac{\deg(u)}{\deg(v)})$ . However, it was shown in [17,35] that RW (after re-weighting, as in Section 2.4) outperforms MHRW for most applications. We therefore restrict our attention to comparing against RW.

**Weighted Random Walk (WRW)** is RW on a weighted graph [4]. At node  $u$ , WRW chooses the edge  $\{u, v\}$  to follow with probability  $P_{u,v}$  proportional to the weight  $w(u, v) \geq 0$  of this edge, *i.e.*,

$$P_{u,v} = \frac{w(u, v)}{\sum_{v' \in \mathcal{N}(u)} w(u, v')}. \quad (8)$$

The stationary distribution of WRW is:

$$\pi^{\text{WRW}}(v) = \frac{w(v)}{\sum_{u \in V} w(u)}. \quad (9)$$

WRW is the basic building block of our design. In the next sections, we show how to choose weights for a specific estimation problem.

**Graph Traversals (BFS, DFS, RDS, ...)** is a family of crawling techniques where no node is sampled more than once. Because traversals introduce a generally unknown bias (see Sec. 7), we do not consider them in this paper.

### 2.4 Correcting the bias

RW, WRW, and WIS all produce biased (nonuniform) node samples. But their bias is known and therefore can be corrected by an appropriate re-weighting of the measured

values. This can be done using the Hansen-Hurwitz estimator [19] as first shown in [39,41] for random walks and also used in [35]. Let every node  $v \in V$  carry a value  $x(v)$ . We can estimate the population total  $x_{\text{tot}} = \sum_v x(v)$  by

$$\hat{x}_{\text{tot}} = \frac{1}{n} \sum_{v \in S} \frac{x(v)}{\pi(v)}, \quad (10)$$

where  $\pi(v)$  is the sampling probability of node  $v$  in the stationary distribution. In practice, we usually know  $\pi(v)$ , and thus  $\hat{x}_{\text{tot}}$ , only up to a constant, *i.e.*, we know the (non-normalized) weights  $w(v)$ . This problem disappears when we estimate the population mean  $x_{\text{av}} = \sum_v x(v)/N$  as

$$\hat{x}_{\text{av}} = \frac{\sum_{v \in S} \frac{x(v)}{\pi(v)}}{\sum_{v \in S} \frac{1}{\pi(v)}} = \frac{\sum_{v \in S} \frac{x(v)}{w(v)}}{\sum_{v \in S} \frac{1}{w(v)}}. \quad (11)$$

For example, for  $x(v) = 1$  if  $\deg(v) = k$  (and  $x(v) = 0$  otherwise),  $\hat{x}_{\text{av}}(k)$  estimates the node degree distribution in  $G$ .

All the results in this paper are presented *after this re-weighting* step, whenever necessary.

## 3. STRATIFIED SAMPLING

In Sec. 1, we argued that in order to compare the median income of residents of China and Vatican we should take 50 random samples from each of these two countries, rather than taking 100 UIS samples from China and Vatican together (or, even worse, from the world's population). This problem naturally arises in the field of survey sampling. The most common solution is *stratified sampling* [12, 28,34], where nodes  $V$  are partitioned into a set  $\mathcal{C}$  of non-overlapping node categories (or "strata"), with  $\bigcup_{C \in \mathcal{C}} C = V$ . Next, we select uniformly at random  $n_i$  nodes from category  $C_i$ . We are free to choose the allocation  $(n_1, n_2, \dots, n_{|\mathcal{C}|})$ , as long as we respect the total budget of samples  $n = \sum_i n_i$ .

Under *proportional allocation* [28] (or "prop") we use  $n_i \propto |C_i|$ , *i.e.*,

$$n_i^{\text{prop}} = |C_i| \cdot n/N. \quad (12)$$

Another possibility is to do an *optimal* allocation (or "opt") that minimizes the variance  $\mathbb{V}$  of our estimator for the specific problem of interest. For example, assume that every node  $v \in V$  carries a value  $x(v)$ , and we may want to estimate the mean of  $x$  in various scenarios, as discussed below.

### 3.1 Examples of Stratified Sampling Problems

#### 3.1.1 Estimating the mean across the entire $V$

A classic application of stratification is to better estimate the population mean  $\mu$ , given several groups (strata) of different properties (*e.g.*, variances). Given  $n_i$  samples from category  $C_i$ , we can estimate the mean  $\mu_i = \frac{1}{|C_i|} \sum_{v \in C_i} x(v)$  over category  $C_i$  by

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{v \in S \cap C_i} x(v) \quad \text{with} \quad \mathbb{V}(\hat{\mu}_i) = \frac{\sigma_i^2}{n_i}, \quad (13)$$

where  $\mathbb{V}(\hat{\mu}_i)$  is the variance of this estimator and  $\sigma_i^2$  is the variance of population  $C_i$ . We can estimate population mean  $\mu$  by a weighted average over all  $\hat{\mu}_i$ s [28], *i.e.*,

$$\hat{\mu} = \sum_i \frac{|C_i|}{N} \cdot \hat{\mu}_i \quad \text{with} \quad \mathbb{V}(\hat{\mu}) = \sum_i \frac{(|C_i|)^2 \cdot \sigma_i^2}{N^2 \cdot n_i}.$$

Under proportional allocation (Eq.(12)), this boils down to  $\mathbb{V}(\hat{\mu}^{\text{prop}}) = \frac{1}{N \cdot n} \sum_i |C_i| \cdot \sigma_i^2$ . However, we can apply Lagrange multipliers to find that  $\mathbb{V}(\hat{\mu})$  is minimized when

$$n_i^{\text{opt}} = \frac{|C_i| \cdot \sigma_i}{\sum_j |C_j| \cdot \sigma_j} \cdot n. \quad (14)$$

This solution is sometimes called ‘Neyman allocation’ [34]. This gives us the variance under optimal allocation  $\mathbb{V}(\hat{\mu}^{\text{opt}}) = \frac{1}{N^2 \cdot n} \left( \sum_i |C_i| \cdot \sigma_i \right)^2$ .

The variances  $\mathbb{V}(\hat{\mu}^{\text{prop}})$  and  $\mathbb{V}(\hat{\mu}^{\text{opt}})$  are measures of the performance of proportional and optimal allocation, respectively. In order to make their practical interpretation easier, we also show how these variances translate into sample lengths. We define as *gain*  $\alpha$  of ‘opt’ over ‘prop’ the number of times ‘prop’ must be longer than ‘opt’ in order to achieve the same variance

$$\text{gain } \alpha = \frac{n^{\text{prop}}}{n^{\text{opt}}}, \text{ subject to } \mathbb{V}^{\text{prop}} = \mathbb{V}^{\text{opt}}.$$

In that case, the gain is

$$\alpha = N \cdot \frac{\sum_i |C_i| \cdot \sigma_i^2}{\left( \sum_i |C_i| \cdot \sigma_i \right)^2} \quad (\geq 1). \quad (15)$$

Notice that this gain does not depend on the sample budget  $n$ . The gain is one of the main metrics we will use in the evaluation sections to assess the efficiency of our technique compared to the random walk.

### 3.1.2 Highest precision for all categories

If we are equally interested in each category, we might want the same (highest possible) precision of estimating  $\mu_i$  for all categories  $C_i$ . In this case, the metric to minimize is  $\mathbb{V}_{\max} = \max_i \{\mathbb{V}(\hat{\mu}_i)\} = \max_i \left\{ \frac{\sigma_i^2}{n_i} \right\}$ . Under proportional allocation, this translates to  $\mathbb{V}_{\max}^{\text{prop}} = \frac{N}{n} \max_i \frac{\sigma_i^2}{|C_i|}$ . But the optimal  $n_i$ , which makes  $\mathbb{V}(\hat{\mu}_i)$  equal for all  $i$ , is

$$n_i^{\text{opt}} = \frac{\sigma_i^2}{\sum_j \sigma_j^2} \cdot n. \quad (16)$$

Consequently,  $\mathbb{V}_{\max}^{\text{opt}} = \frac{\sum_i \sigma_i^2}{n}$ , which leads to gain

$$\alpha = \frac{\max_i \left\{ \frac{N}{|C_i|} \sigma_i^2 \right\}}{\sum_i \sigma_i^2} \quad (\geq 1). \quad (17)$$

### 3.1.3 Smallest sum of variances across categories

Even if we are interested in all categories, an alternative objective is to maximize the *average* precision of category pair comparisons (see Sec. 5A.13 in [12]), which is equivalent to minimizing the sum  $\mathbb{V}_{\Sigma} = \sum_i \mathbb{V}(\hat{\mu}_i) = \sum_i \frac{\sigma_i^2}{n_i}$ . In this case, proportional allocation achieves  $\mathbb{V}_{\Sigma}^{\text{prop}} = \frac{N}{n} \sum_i \frac{\sigma_i^2}{|C_i|}$ , while, using Lagrange multipliers we get

$$n_i^{\text{opt}} = \frac{\sigma_i}{\sum_j \sigma_j} \cdot n \quad \text{and} \quad \mathbb{V}_{\Sigma}^{\text{opt}} = \frac{(\sum_i \sigma_i)^2}{n}, \quad (18)$$

which leads to gain

$$\alpha = \frac{\sum_i \frac{N}{|C_i|} \sigma_i^2}{(\sum_i \sigma_i)^2} \quad (\geq 1). \quad (19)$$

### 3.1.4 Relative sizes of node categories

Stratified sampling assumes that we know the sizes  $|C_i|$  of node categories. In some applications, however, these sizes are unknown and among the values we need to estimate as well (*e.g.*, by using UIS or WIS). We show in Appendix C (for  $|\mathcal{C}| = 2$ ) that the optimal sample allocation and the corresponding gain  $\alpha$  of WIS over UIS are respectively

$$n_i^{\text{WIS}} = \frac{1}{|\mathcal{C}|} \cdot n \quad \text{and} \quad \alpha = \frac{N^2}{4|C_1| \cdot |C_2|}. \quad (20)$$

### 3.1.5 Irrelevant category $C_{\Theta}$ (aggregated)

In many practical cases, we may want to measure some (but not all) node categories. *E.g.*, in Fig. 1, we are interested in blue and black nodes, but not in white ones. Similarly, in our Facebook study in Section 6 we are only interested in self-declared college students, which accounts for only 3.5% of all users. We group all categories not covered by our measurement objective as a single *irrelevant category*  $C_{\Theta} \in \mathcal{C}$ , and we set  $n_{\Theta}^{\text{opt}} = 0$ . In contrast,  $n_{\Theta}^{\text{prop}} = |C_{\Theta}| \cdot n/N$ . As a result, under ‘opt’ we have  $N/(N - |C_{\Theta}|)$  times more useful samples than under ‘prop’. Now, if we allocate optimally all these useful samples between the relevant categories  $\mathcal{C} \setminus \{C_{\Theta}\}$ , the gain  $\alpha$  becomes

$$\alpha = \frac{N}{N - |C_{\Theta}|} \cdot \alpha(\mathcal{C} \setminus \{C_{\Theta}\}), \quad (21)$$

where  $\alpha(\mathcal{C} \setminus \{C_{\Theta}\})$  is the gain (15), (17), (19) or (20), depending on the metric, calculated only within categories  $\mathcal{C} \setminus \{C_{\Theta}\}$ .

In other words, gain  $\alpha$  is now composed of two factors: (i) gain in avoiding irrelevant categories, and (ii) gain in optimal allocation of samples among the relevant categories.

### 3.1.6 Practical Guideline

Let us look at the optimal weights in the above scenarios, when all  $\sigma_i = \sigma$  are the same. This is a reasonable working assumption in many practical settings, since we typically do not have prior estimates of  $\sigma_i$ . With this simplification, Eq.(14) becomes

$$n_i^{\text{opt}} = \frac{|C_i|}{N} \cdot n = n_i^{\text{prop}}.$$

In contrast, Eq.(16), Eq.(18) and Eq.(20) get simplified to

$$n_i^{\text{opt}} = \frac{1}{|\mathcal{C}|} \cdot n.$$

In conclusion, if we are interested in comparing the node categories with respect to some properties (*e.g.*, average node degree, category size), rather than estimating a property across the entire population, we should take an *equal number of samples from every relevant category*.

## 4. EDGE WEIGHT SETTING UNDER WRW

In the previous section, we studied the optimal sample allocation under (independence) stratified sampling. However, independence node sampling is typically impossible in large online graphs, while crawling the graph is a natural, available exploration primitive. In this section, we show how to perform a weighted random walk (WRW) which approximates the stratified sampling of the previous section. We can formulate the general problem as follows:

Given a measurement objective, error metric and sampling budget  $|S|=n$ , set the edge weights in graph  $G$  such that the WRW measurement error is minimized.

Although we are able to solve this problem analytically for some specific and fully known topologies, it is not obvious how to address it in general, especially under a limited knowledge of  $G$ . Instead, in this paper, we propose S-WRW, a heuristic to set the edge weights. S-WRW starts from a solution optimal under WIS, and takes into account practical issues that arise in graph exploration. Once the weights are set, we simply perform WRW as described in Section 2.3 and collect samples.

## 4.1 Preliminaries

### 4.1.1 Category-level granularity

One can think of the problem in two levels of granularity: the original graph  $G = (V, E)$  and the *category graph*  $G^C = (\mathcal{C}, E^C)$ . In  $G^C$ , nodes represent categories, and every undirected edge  $\{C_1, C_2\} \in E^C$  represents the corresponding non-empty set of edges  $E_{C_1, C_2} \subset E$  in the original graph  $G$ , *i.e.*,

$$E_{C_1, C_2} = \{ \{u, v\} \in E : u \in C_1 \text{ and } v \in C_2 \} \neq \emptyset.$$

In our approach, we move from the finer granularity of  $G$  to the coarser granularity of  $G^C$ . This means that we are interested in collecting, say,  $n_i$  samples from category  $C_i$ , but we do not control how these  $n_i$  nodes are collected (*i.e.*, with what individual sampling probabilities).

The rationale for that simplification is twofold. From a theoretical point of view, categories are exactly the properties of interest in the estimation problems we consider. From a practical point of view, it is relatively easy to obtain or infer information about categories, as we show *e.g.*, in Sec. 4.2.1.

### 4.1.2 Stratification in expectation

Ideally, we would like to enforce strictly stratified sampling. However, when we use crawling instead of independence sampling, sampling exactly  $n_i$  nodes from category  $C_i$  (and no other nodes) is possible only by discarding observations. It is thus more natural to frame the problem in terms of the probability mass placed on each category in equilibrium. This can be achieved by making the weight  $w(C_i)$  of each category proportional to the desired number  $n_i$  of samples, *i.e.*,

$$w(C_i) \propto n_i. \quad (22)$$

As a result, we draw  $n_i$  samples from  $C_i$  in expectation.

### 4.1.3 Main guideline

As the main guideline, S-WRW tries to realize the category weights  $w^{\text{WIS}}(C_i)$  that are optimal under WIS. There are many edge weight settings in  $G$  that achieve  $w^{\text{WIS}}(C_i)$ . In our implementation, we observe that  $\text{vol}(C_i)$  counts the number of edges incident on nodes of  $C_i$ . Consequently, if for every category  $C_i$  we set in  $G$  the weights of all edges incident on nodes in  $C_i$  to

$$w_e(C_i) = \frac{w^{\text{WIS}}(C_i)}{\text{vol}(C_i)}. \quad (23)$$

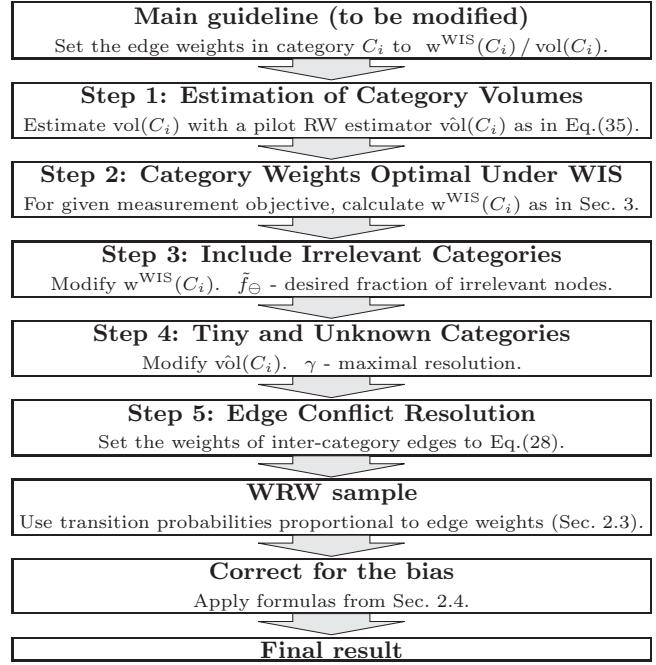


Figure 2: Overview of our approach.

then weight  $w^{\text{WIS}}(C_i)$  are achieved.<sup>2</sup> This simple observation is central to the S-WRW heuristic.

In order to apply Eq.(23), we first have to calculate or estimate its terms  $\text{vol}(C_i)$  and  $w^{\text{WIS}}(C_i)$ .<sup>3</sup> Below, we show how to do it in Step 1 and 2, respectively. Next, in Steps 3-5, we show how to modify these terms to account for practical problems arising mainly from the underlying graph structure.

## 4.2 Our practical solution: S-WRW

### 4.2.1 Step 1: Estimation of Category Volumes

In general, we have no prior information about  $G$  or  $G^C$ . Fortunately, it is easy and inexpensive estimate the relative category volumes  $f_i^{\text{vol}}$  which is the first piece of information we need in Eq.(23) (see footnote 3). Indeed, it is enough to run a relatively short pilot RW, and plug the collected sample  $S$  in Eq.(35) derived in Appendix B, as follows

$$\hat{f}_i^{\text{vol}} = \frac{1}{n} \sum_{u \in S} \left( \frac{1}{\deg(u)} \sum_{v \in \mathcal{N}(u)} 1_{\{v \in C_i\}} \right).$$

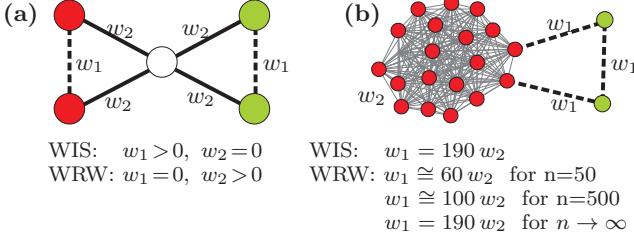
### 4.2.2 Step 2: Category Weights Optimal Under WIS

In order to find the optimal WIS category weights  $w^{\text{WIS}}(C_i)$  in Eq.(23), we first calculate  $n_i^{\text{opt}}$  as shown, under various scenarios, in Sec. 3. Next, we plug the resulting  $n_i^{\text{opt}}$  in Eq.(22), *e.g.*, by setting  $w^{\text{WIS}}(C_i) = n_i^{\text{opt}}$ .

### 4.2.3 Step 3: Irrelevant Categories

<sup>2</sup>There exist many other edge weight assignments that lead to  $w^{\text{WIS}}(C_i)$ . Eq.(23) has the advantage of distributing the weights evenly across all  $\text{vol}(C_i)$  edges.

<sup>3</sup>In fact, we need to know  $w_e(C_i)$  in Eq.(23) only *up to a constant factor*, because these factors cancel out in the calculation of transition probabilities of WRW in Eq.(8). Consequently, the same applies to  $\text{vol}(C_i)$  and  $w^{\text{WIS}}(C_i)$ .



**Figure 3: Optimal edge weights: WIS vs WRW.** The objective is to compare the sizes of red (dark) and green (light) categories.

**Problem: Potentially poor or no convergence.** Consider the toy example in Fig. 3(a). We are interested in finding the relative sizes of red (dark) and green (light) categories. The white node in the middle is irrelevant for our measurement objective. Due to symmetry, we distinguish between two types of edges with weights  $w_1$  and  $w_2$ . Under WIS, Eq.(20) gives us the optimal weights  $w_1 > 0$  and  $w_2 = 0$ , *i.e.*, WIS samples every non-white node with the same probability and never samples the white one. However, under WRW with these weights, relevant nodes get disconnected into two components and WRW does not converge. We observed a similar problem in Fig. 1.

**Guideline: Occasionally visit irrelevant nodes.** We show in Appendix D that the optimal WRW weights in Fig. 3(a) are  $w_1 = 0$  and  $w_2 > 0$ . In that case, half of the samples are due to visits in the white (irrelevant) node. In other words, WRW may benefit from allocating small weight  $w(C_\Theta) > 0$  to category  $C_\Theta$  that groups all (if any) categories irrelevant to our estimation. The intuition is that irrelevant nodes may not contribute to estimation but may be needed for connectivity or fast mixing.

**Implementation in S-WRW.** In S-WRW, we achieve this goal by replacing  $w^{\text{WIS}}(C_i)$  with

$$\tilde{w}^{\text{WIS}}(C_i) = \begin{cases} w^{\text{WIS}}(C_i) & \text{if } C_i \neq C_\Theta \\ \tilde{f}_\Theta \cdot \sum_{C \neq C_\Theta} w^{\text{WIS}}(C) & \text{if } C_i = C_\Theta. \end{cases} \quad (24)$$

The parameter  $0 \leq \tilde{f}_\Theta \ll 1$  controls the desired fraction of visits in  $C_\Theta$ .

#### 4.2.4 Step 4: Tiny and Unknown Categories

**Problem: “black holes”.** Every optical system has a fundamental magnification limit due to diffraction and our “graph magnifying glass” is no exception. Consider the toy graph in Fig. 3(b): it consists of a big clique  $C_{\text{big}}$  of 20 red nodes with edge weights  $w_2$ , and a green category  $C_{\text{tiny}}$  with two nodes only and edge weights  $w_1$ . In Sec. 3.1.4, we saw that WIS optimally estimates the relative sizes of red and green categories for  $w(C_{\text{big}}) = w(C_{\text{tiny}})$ , *i.e.*, for  $w_1 = 190 w_2$ . However, for such large values of  $w_1$ , the two green nodes behave as a sink (or a “black hole”) for a WRW of finite length, thus increasing the variance of the category size estimation.

**Guideline: limit edge weights.** In other words, although WIS suggests to over-sample small categories, WRW should “under-over-sample” very small categories to avoid black holes. For example, in Fig. 3(b)  $w_1 \simeq 60 w_2$  ( $\ll 190 w_2$ ) is optimal for WRW of length  $n=50$  (simulation results).

**Implementation in S-WRW.** In S-WRW, we achieve this

goal by replacing  $\text{vol}(C_i)$  in Eq.(23) with

$$\tilde{\text{vol}}(C) = \max \{ \text{vol}(C), \text{vol}_{\min} \}, \quad \text{where} \quad (25)$$

$$\text{vol}_{\min} = \frac{1}{\gamma} \cdot \max_{C \neq C_\Theta} \{ \text{vol}(C) \}. \quad (26)$$

Moreover, this formulation takes care of every category  $C$  that was not discovered by the pilot RW in Sec. 4.2.1, by setting  $\tilde{\text{vol}}(C) = \text{vol}_{\min}$ .

#### 4.2.5 Step 5: Edge Conflict Resolution

**Problem: Conflicting desired edge weights.** With the above modifications, our target edge weights defined in Eq.(23) can be rewritten as

$$\tilde{w}_e(C_i) = \frac{\tilde{w}^{\text{WIS}}(C_i)}{\tilde{\text{vol}}(C_i)}. \quad (27)$$

We can directly set the weight  $w(u, v) = \tilde{w}_e(C(u)) = \tilde{w}_e(C(v))$  for every intra-category edge  $\{u, v\}$ . However, for every inter-category edge, we usually have “conflicting” weights  $\tilde{w}_e(C(u)) \neq \tilde{w}_e(C(v))$  desired at the two ends of the edge.

**Guideline: prefer inter-category edges.** There are several possible edge weight assignments that achieve the desired category node weights. High weights on intra-category edges and small weights on inter-category edges result in WRW staying in small categories  $C_{\text{tiny}}$  for a long time. In order to improve the mixing time, we should do exactly the opposite, *i.e.*, assign relatively high weights to inter-category edges (connecting relevant categories). As a result, WRW will enter  $C_{\text{tiny}}$  more often, but will stay there for a short time. This intuition is motivated by Monte Carlo variance reduction techniques such as the use of *antithetic variates* [15], which seek to induce negative correlation between consecutive draws so as to reduce the variance of the resulting estimator.

**Implementation in S-WRW.** We choose to assign an edge weight  $\tilde{w}_e$  that is in between these two values  $\tilde{w}_e(C(u))$  and  $\tilde{w}_e(C(v))$ . We considered several candidate such assignments. We may take the *arithmetic* or *geometric* mean of the conflicting weights, which we denote by  $w^{\text{ar}}(u, v)$  and  $w^{\text{ge}}(u, v)$ , respectively. We may also use the *maximum* of the two values,  $w^{\text{max}}(u, v)$ , which should improve mixing according to the discussion above. However,  $w^{\text{max}}(u, v)$  alone would also add high weight to irrelevant nodes  $C_\Theta$  (possibly far beyond  $\tilde{f}_\Theta$ ). To avoid this undesired effect, we distinguish between the two cases by defining a hybrid solution:

$$w^{\text{hy}}(u, v) = \begin{cases} w^{\text{ge}}(u, v) & \text{if } C_\Theta \in \{C(u), C(v)\} \\ w^{\text{max}}(u, v) & \text{otherwise.} \end{cases} \quad (28)$$

This hybrid edge assignment was the one we found to work best in practice - see Section 6.

### 4.3 Discussion

#### 4.3.1 Information needed about the neighbors

In the pilot RW (Sec. 4.2.1) as well as in the main WRW, we assume that by sampling a node  $v$  we also learn the category (but not degree) of each of its neighbors  $u \in \mathcal{N}(v)$ . Fortunately, such information is often available in most online graphs at no additional cost, especially when scraping html pages (as we do). For example, when sampling colleges

in Facebook (Sec. 6), we use the college membership information of all  $v$ 's neighbors, which, in Facebook, is available at  $v$  together with the friends list.

### 4.3.2 Cost of pilot RW

The pilot RW volume estimator described in Sec. 4.2.1 considers the categories not only of the sampled nodes, but also of their neighbors. As a result, it achieves high efficiency, as we show in simulations (Sec. 5.3.1) and Facebook measurements (Sec. 6.1). Given that, and high robustness of S-WRW to estimation errors (see Sec. 5.3.5), pilot RW should be only a small fraction of the later WRW (*e.g.*, 6.5% in our Facebook measurements in Sec. 6).

### 4.3.3 Setting the parameters

S-WRW sets the edge weights trying to achieve roughly  $w^{\text{WIS}}(C_i)$  as the main goal. We slightly shape  $w^{\text{WIS}}(C_i)$  to avoid black holes and improve mixing, which is controlled by two natural and easy-to-interpret parameters,  $\tilde{f}_\Theta$  and  $\gamma$ .

**Irrelevant nodes visits  $\tilde{f}_\Theta$ .** The parameter  $0 \leq \tilde{f}_\Theta \ll 1$  controls the desired fraction of visits in  $C_\Theta$ . When setting  $\tilde{f}_\Theta$ , we should exploit the information provided by the pilot crawl. If the relevant categories appear poorly interconnected and often separated by irrelevant nodes, we should set  $\tilde{f}_\Theta$  relatively high. We have seen an extreme case in Fig. 3(a), with disconnected relevant categories and optimal  $\tilde{f}_\Theta = 0.5$ . In contrast, when the relevant categories are strongly interconnected, we should use much smaller  $\tilde{f}_\Theta$ . However, because we can never be sure that the graph induced on relevant nodes is connected, we recommend always using  $\tilde{f}_\Theta > 0$ . For example, when measuring Facebook in Sec. 6, we set  $\tilde{f}_\Theta = 1\%$ .

**Maximal resolution  $\gamma$ .** The parameter  $\gamma \geq 1$  can be interpreted as the maximal resolution of our ‘‘graph magnifying glass’’, with respect to the largest relevant category  $C_{\text{big}}$ . S-WRW will typically sample well all categories that are less than  $\gamma$  times smaller than  $C_{\text{big}}$ ; all categories smaller than that are relatively undersampled (see Sec. 6.2.4). In the extreme case, for  $\gamma \rightarrow \infty$ , S-WRW tries to cover every category, no matter how small, which may cause the ‘‘black hole’’ problem discussed in Sec. 4.2.4. In the other extreme, for  $\gamma = 1$  (and identical  $w^{\text{WIS}}(C_i)$  for all categories, including  $C_\Theta$ ), S-WRW reduces to RW. We recommend always setting  $1 < \gamma < \infty$ . Ideally, we know  $|C_{\text{smallest}}|$  - the smallest category size that is still relevant to us. In that case we should set  $\gamma = |C_{\text{big}}|/|C_{\text{smallest}}|$ .<sup>4</sup> For example, in Sec. 6 the categories are US colleges; we set  $\gamma = 1000$ , because colleges with size smaller than 1/1000th of the largest one (*i.e.*, with a few tens of students) seem irrelevant to our measurement. As another rule of thumb, we should try to set smaller  $\gamma$  for relatively small sample sizes and in graphs with tight community structure (see Sec. 5.3.5).

### 4.3.4 Conservative approach

Note that a reasonable setting of these parameters (*i.e.*,  $\tilde{f}_\Theta > 0$  and  $1 < \gamma < \infty$ , and any conflict resolution discussed in the paper), increases the weights of large categories (including  $C_\Theta$ ) and decreases the weight of small categories,

<sup>4</sup>Strictly speaking,  $\gamma$  is related to volumes  $\text{vol}(C_i)$  rather than sizes  $|C_i|$ . They are equivalent when category volume is proportional to its size, which is often the case, and is the central assumption in the ‘‘scale-up method’’ [9].

compared to  $w^{\text{WIS}}(C_i)$ . This makes S-WRW allocate category weights between the two extremes: RW and WIS. Consequently, S-WRW can be considered *conservative* (with respect to WIS).

### 4.3.5 S-WRW is unbiased

It is also important to note that because the collected WRW sample is eventually corrected with the actual sampling weights as described in Sec. 2.4, S-WRW estimation process is *unbiased*, regardless of the choice of weights (so long as convergence is attained). In contrast, suboptimal weights (*e.g.*, due to estimation error of  $\hat{f}_C^{\text{vol}}$ ) can increase WRW mixing time, and/or the *variance* of the resulting estimator. However, our simulations and empirical experiments on Facebook (see Sec. 5 and 6) show that S-WRW is very robust to suboptimal choice of weights.

## 5. SIMULATION RESULTS

The gain of our approach compared to RW comes from two main factors. First, S-WRW avoids, to a large extent or completely, the nodes in  $C_\Theta$  that are irrelevant to our measurement. This fact alone can bring an arbitrarily large improvement ( $\frac{N}{N-|C_\Theta|}$  under WIS), especially when  $C_\Theta$  is large compared to  $N$ . We demonstrate this in the Facebook measurements in Section 6. Second, we can better allocate samples among the relevant categories. This factor is observable in our Facebook measurements as well, but it is more difficult to evaluate due to the lack of ground-truth therein. In this section, we evaluate the optimal allocation gain in a controlled simulation and we demonstrate some key insights.

### 5.1 Setup

We consider a graph  $G$  with 101K nodes and 505.5K edges organized in two densely (and randomly) connected communities<sup>5</sup> as shown in Fig. 4(h).

The nodes in  $G$  are partitioned into two node categories:  $C_{\text{tiny}}$  with 1K nodes (dark red), and  $C_{\text{big}}$  with 100K nodes (light yellow). We consider two extreme scenarios of such a partition. The ‘‘random’’ scenario is purely random, as shown in Fig. 4(a). In contrast, under ‘‘clustered’’, categories  $C_{\text{tiny}}$  and  $C_{\text{big}}$  coincide with the existing communities in  $G$ , as shown in Fig. 4(h). It is arguably the worst case scenario for graph sampling by exploration.

We fix the edge weights of all internal edges in  $C_{\text{big}}$  to 1. All the remaining edges, *i.e.*, all edges incident on nodes in category  $C_{\text{tiny}}$ , have weight  $w$  each, where  $w \geq 1$  is a parameter. Note that this is equivalent to setting  $\tilde{w}_e(C_{\text{big}}) = 1$ ,  $\tilde{w}_e(C_{\text{tiny}}) = w$ , and ‘‘max’’ or ‘‘hybrid’’ conflict resolution.

### 5.2 Measurement objective and error metric

We are mainly interested in measuring the relative sizes  $f_{\text{tiny}}$  and  $f_{\text{big}}$  of categories  $C_{\text{tiny}}$  and  $C_{\text{big}}$ , respectively.

We use Normalized Root Mean Square Error (NRMSE) to assess the estimation error, defined as [37]:

$$\text{NRMSE}(\hat{x}) = \frac{\sqrt{\mathbb{E}[(\hat{x} - x)^2]}}{x}, \quad (29)$$

where  $x$  is the real value and  $\hat{x}$  is the estimated one.

<sup>5</sup>The term ‘‘community’’ refers to cluster and is defined purely based on topology. The term ‘‘category’’ is a property of a node and is independent of topology.

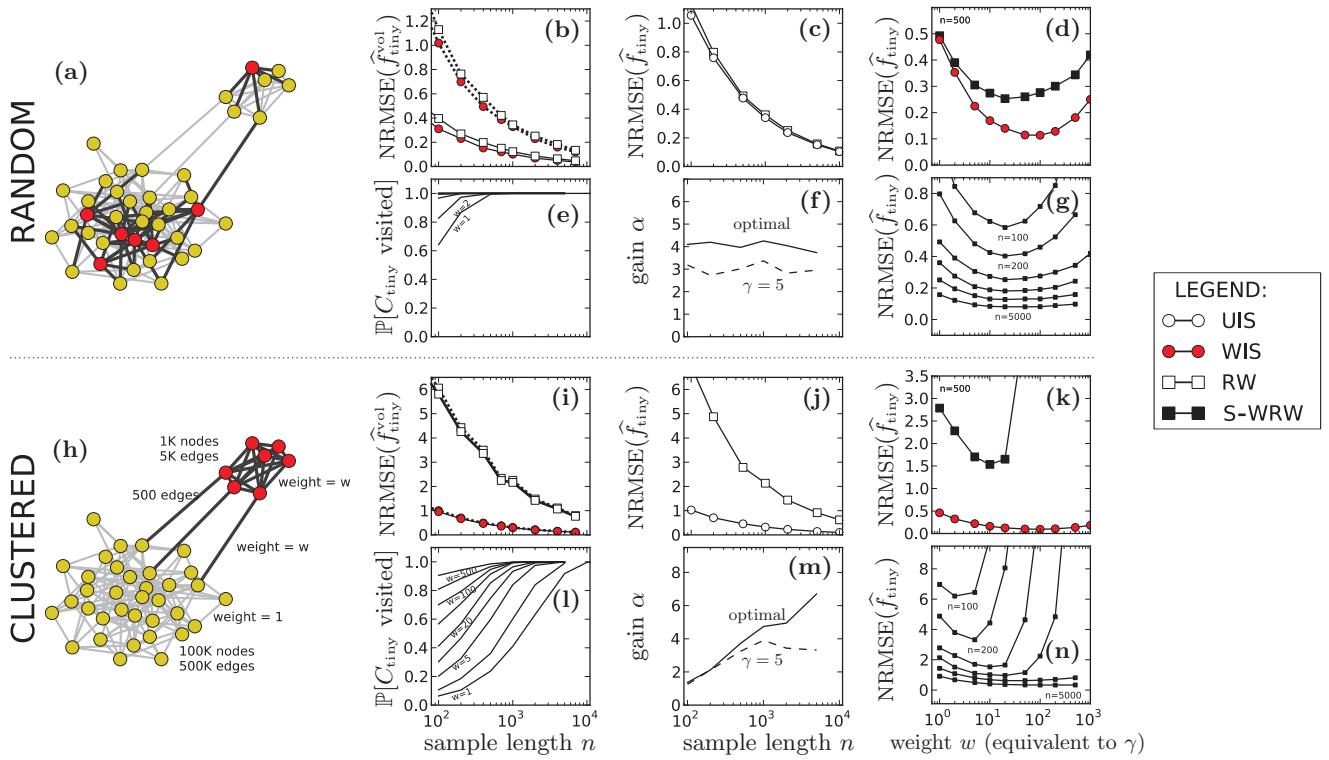


Figure 4: RW and S-WRW under two scenarios: Random (a-g) and Clustered (h-n). In (b,i), we show error of two volume estimators: naive Eq.(32) (dotted) and neighbor-based Eq.(35) (plain). Next, we show error of size estimator as a function of  $n$  (c,j) and  $w$  (d,g,k,n); in the latter, UIS and RW correspond to WIS and S-WRW for  $w=1$ . In (e,l), we show the empirical probability that S-WRW visits  $C_{\text{tiny}}$  at least once. Finally, (f,m) is gain  $\alpha$  of S-WRW over RW under the optimal choice of  $w$  (plain), and for fixed  $\gamma=w=5$  (dashed).

## 5.3 Results

### 5.3.1 Estimating volumes is usually cheap

The first step in S-WRW is obtaining category volume estimates  $\hat{f}_i^{\text{vol}}$ . We achieve it by running a short pilot RW and applying the estimator Eq.(35). We show NRMSE( $\hat{f}_{\text{tiny}}^{\text{vol}}$ ) as plain curves in Fig. 4(b). This estimator takes advantage of the knowledge of the categories of the neighboring nodes, which makes it much more efficient than the naive estimator Eq.(32) shown by dashed curves. Moreover, the advantage of Eq.(35) over Eq.(32) grows with the graph density, and the skewness of its degree distribution (not shown here).

Note that under ‘random’, RW and WIS (with the sampling probabilities of RW) are almost equally efficient. However, on the other extreme, *i.e.*, under the ‘clustered’ scenario, the performance of RW becomes much worse and the advantage of Eq.(35) over Eq.(32) diminishes. This is because essentially all friends of a node from category  $C_i$  are in  $C_i$  too, which reduces formula Eq.(35) to Eq.(32). Nevertheless, we show later in Sec. 5.3.5 that even severalfold volume estimation errors are likely not to affect significantly the results.

### 5.3.2 Visiting the tiny category

Fig. 4(e,l) presents the empirical probability  $\mathbb{P}[C_{\text{tiny}} \text{ visited}]$  that our walk visits at least one node from  $C_{\text{tiny}}$ . Of course, this probability grows with the sample length. However, the choice of weight  $w$  also helps in it. Indeed, WRW with  $w > 1$  is more likely to visit  $C_{\text{tiny}}$  than RW ( $w = 1$ , bottom line).

This demonstrates the first advantage of introducing edge weights and WRW.

### 5.3.3 Optimal $w$ and $\gamma$

Let us now focus on the estimation error as a function of  $w$ , shown in Fig. 4(d,k). Interestingly, this error does not drop monotonically with  $w$  but follows a ‘U’ shaped function with a clear optimal value  $w^{\text{opt}}$ .

Under WIS, we have  $w^{\text{opt}} \simeq 100$ , which confirms our findings in Sec. 3.1.4. Indeed, according to Eq.(20), we need the same number of samples from the two categories, and thus  $w^{\text{WIS}}(C_{\text{tiny}}) = w^{\text{WIS}}(C_{\text{big}})$  (by Eq.(22)). By plugging this and  $\text{vol}(C_{\text{big}}) = 100 \cdot \text{vol}(C_{\text{tiny}})$  to Eq.(23), we finally obtain the WIS-optimal edge weights in  $C_{\text{tiny}}$ , *i.e.*,  $w^{\text{opt}} = w_e(C_{\text{tiny}}) = 100 \cdot w_e(C_{\text{big}}) = 100$ .<sup>6</sup>

In contrast, WRW is optimized for  $w < 100$ . For the sample length  $n = 500$  as in Fig. 4(d,k), the error is minimized already for  $w^{\text{opt}} \simeq 20$  and increases for higher weights. This demonstrates the “black hole” effect discussed in Sec. 4.2.4. It is much more pronounced in the ‘clustered’ scenario, confirming our intuition that black-holes become a problem only in the presence of relatively isolated, tight communities. Of course, the black hole effect diminishes with the sample length  $n$  (and completely vanishes for  $n \rightarrow \infty$ ), which can be observed in Fig. 4(g,n), especially in (n).

In other words, the optimal assignment of edge weights (in relevant categories) under WRW lies somewhere between

<sup>6</sup>For simplicity, we ignored in this calculation the conflicts on the 500 edges between  $C_{\text{big}}$  and  $C_{\text{tiny}}$ .

RW (all weights equal) and WIS. In S-WRW, we control it by parameter  $\gamma$ . In this example, we have  $\gamma \equiv w$  for  $\gamma \leq 100$ . Indeed, by combining Eq.(23), Eq.(25), Eq.(26),  $w^{\text{WIS}}(C_{\text{tiny}}) = w^{\text{WIS}}(C_{\text{big}})$ , we obtain

$$\begin{aligned} w &= \frac{w}{1} = \frac{w_e(C_{\text{tiny}})}{w_e(C_{\text{big}})} = \frac{w^{\text{WIS}}(C_{\text{tiny}})/\tilde{\text{vol}}(C_{\text{tiny}})}{w^{\text{WIS}}(C_{\text{big}})/\tilde{\text{vol}}(C_{\text{big}})} \\ &= \frac{\tilde{\text{vol}}(C_{\text{big}})}{\tilde{\text{vol}}(C_{\text{tiny}})} = \frac{\text{vol}(C_{\text{big}})}{\frac{1}{\gamma}\text{vol}(C_{\text{big}})} = \gamma. \end{aligned}$$

Consequently, the optimal setting of  $\gamma$  is the same as  $w^{\text{opt}}$  discussed above.

### 5.3.4 Gain $\alpha$

The gain  $\alpha$  of WIS over UIS is given by Eq.(20). In this case, we have  $\alpha = (101K)^2 \cdot (4 \cdot 1K \cdot 100K)^{-1} \simeq 25$ . Indeed, WIS with  $n=500$  samples shown in Fig. 4(d) achieves NRMSE  $\simeq 0.1$ , which is the same as UIS of about  $\alpha = 25$  times more samples (see Fig. 4(c)).

This gain due to stratification is smaller for sampling by exploration: a 500-hop-long WRW with  $w \simeq 20$  yields the same error NRMSE  $\simeq 0.3$  as a 2000-hop-long RW. This means that WRW reduces the sampling cost by a factor of  $\alpha \simeq 4$ . Fig. 4(f) shows that this gain does not vary much with the sampling length. Under ‘clustered’, both RW and WRW perform much worse. Nevertheless, Fig. 4(m) shows that also in this scenario WRW may significantly reduce the sampling cost, especially for longer samples.

It is worth noting that WRW can sometimes significantly outperform UIS. This is the case in Fig. 4(d), where UIS is equivalent to WIS with  $w = 1$ . Because no walk can mix faster than UIS (that is independent and thus has perfect mixing), improving the mixing time alone [5,10,37,38] cannot achieve the potential gains of stratification, in general.

So far we focused on the smaller set  $C_{\text{tiny}}$  only. When estimating the size of  $C_{\text{big}}$ , all errors are much smaller, but we observe similar gain  $\alpha$ .

### 5.3.5 Robustness to $\gamma$ and volume estimation

The gain  $\alpha$  shown above is calculated for the optimal choice of  $w$ , or, equivalently,  $\gamma$ . Of course, in practice it might be impossible to obtain this value. Fortunately, S-WRW is relatively robust to the choice of parameters. The dashed lines in Fig. 4(f,m) are calculated for  $\gamma$  fixed to  $\gamma=5$ , rather than optimized. Note that this value is often drastically smaller than the optimal one (e.g.,  $w^{\text{opt}} \simeq 50$  for  $n = 5000$ ). Nevertheless, although the performance somewhat drops, S-WRW still reduces the sampling cost about three-fold.

This observation also addresses potential concerns one might have regarding the category volume estimation error (see Sec. 4.2.1). Indeed, setting  $\gamma=5$  means that every category  $C_i$  with volume estimated at  $\tilde{\text{vol}}(C_i) \leq \frac{1}{5}\text{vol}(C_{\text{big}})$  is treated the same. In Fig. 4(f), the volume of  $C_{\text{tiny}}$  would have to be overestimated by more than 20 times in order to affect the edge weight setting and thus the results. We have seen in Sec. 5.3.1 that this is very unlikely, even under smallest sample lengths and most adversarial scenarios.

## 5.4 Summary

WRW brings two types of benefits (i) avoiding irrelevant nodes  $C_{\ominus}$  and (ii) carefully allocating samples between relevant categories of different sizes. Even when  $C_{\ominus} = \emptyset$ , WRW

can still reduce the sampling cost by 75%. This second benefit is more difficult to achieve when the categories form strong and tight communities, which leads to the ‘black hole’ effect. We should then choose smaller, more conservative values of  $\gamma$  in S-WRW, which translate into smaller  $w$  in our example. In contrast, under a looser community structure this problem disappears and WRW is closer to WIS.

## 6. IMPLEMENTATION IN FACEBOOK

As a concrete application, we apply S-WRW to measure the Facebook social graph, which is our motivating and canonical example. We also note that it is an undirected and can also be considered a static graph, for all practical purposes in this study.<sup>7</sup> In Facebook, every user may declare herself a member of a college<sup>8</sup> he/she attends. This membership information is publicly available by default and allows us to answer some interesting questions. For example, how do the college networks (or ‘colleges’ for short) compare with respect to their sizes? What is the college-to-college friendship graph? In order to answer these questions, we have to collect many college user samples, preferably evenly distributed between colleges. This is the main goal of this section.

### 6.1 Measurement Setup

By default, every Facebook user can see the basic information on any other user, including the name, photo, and a list of friends together with their college memberships (if any). We developed a high performance multi-threaded crawler to explore Facebook’s social graph by scraping this web interface.

To make informed decision for the parameters of S-WRW, we first ran a short pilot RW (see Sec. 4.2.1) with a total of  $65K$  samples (which is only 6.5% of the length of the main S-WRW sample). Although our pilot walk visited only 2000 colleges, it estimated the relative volumes  $f_i^{\text{vol}}$  for about 9500 colleges discovered among friends of sampled users, as discussed in Sec. 4.3.2. In Fig. 6(a), we show that the neighbor-based estimator Eq.(35) greatly outperforms the naive estimator Eq.(32). These volumes cover several decades. Because colleges with only a few tens of users are not of our interest, we set the maximal resolution to  $\gamma=1000$  (see the discussion in Sec. 4.3.3). Finally, because the college students looked very well interconnected in our pilot RW, we set the desired fraction of irrelevant nodes to a small number  $\tilde{f}_{\ominus}=1\%$ .

In the main measurement phase, we collected three S-WRW crawls, each with different edge weight conflict resolution (hybrid, geometric, and arithmetic), and one simple RW crawl as a baseline comparison (Table 1). For each crawl type we collected 1 million *unique* users. Some of them are sampled multiple times (at no additional bandwidth cost), which results in higher total number of samples in the second row of Table 1. Our crawls were performed on Oct. 16-19 2010, and are available at [1].

<sup>7</sup>The Facebook characteristics do change but in time scales much longer than the 3-day duration of our crawls. Websites such as Facebook statistics, Alexa etc show that the number of Facebook users is growing with rate 0.1-0.2% per day.

<sup>8</sup>There also exist categories other than colleges, namely ‘work’ and ‘high school’. Facebook requires a valid category-specific email for verification.

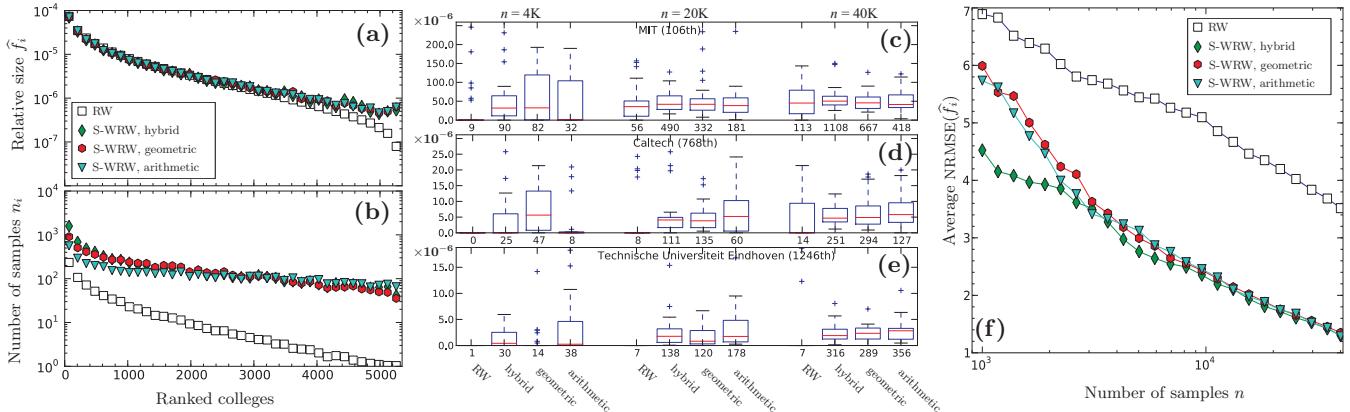


Figure 5: 5331 colleges discovered and ranked by RW. (a) Estimated relative college sizes  $\hat{f}_i$ . (b) Absolute number of user samples per college. (c-e) 25 estimates of size  $\hat{f}_i$  for three different colleges and sample lengths  $n$ . (f) Average NRMSE of college size estimation. Results in (a,b,f) are binned.

	RW	S-WRW		
		Hybrid	Geometric	Arithmetic
Unique samples	1,000K	1,000K	1,000K	1,000K
Total samples	1,016K	1,263K	1,228K	1,237K
College samples	9%	86%	79%	58%
Unique Colleges	5,331	9,014	8,994	10,439

Table 1: Overview of collected Facebook datasets.

## 6.2 Results: RW vs. S-WRW

### 6.2.1 Avoiding irrelevant categories

Only 9% of the RW’s samples come from colleges, which means that the vast majority of sampling effort is wasted. In contrast, the S-WRW crawls achieved 6-10 better efficiency, collecting 86% (hybrid), 79% (geometric) and 58% (arithmetic) samples from colleges. Note that these values are significantly lower than the target 99% suggested by our choice of  $\tilde{f}_\Theta = 1\%$ , and that S-WRW hybrid reaches the highest number. This is in agreement with our discussion in Sec. 4.2.5. Finally, we also note that S-WRW crawls discovered 1.6 – 1.9 times more unique colleges than RW.

It might seem surprising that RW samples colleges in 9% of cases while only 3.5% of Facebook users belong to colleges. This can be explained by looking at the last rows of Table 1. Indeed, the college users have on average three times more Facebook friends than average users, and therefore they attract RW approximately three times more often.

### 6.2.2 Stratification

The advantage of S-WRW over RW does not lie exclusively in avoiding the nodes in the irrelevant category  $C_\Theta$ . S-WRW can also over-sample small categories (here colleges) at the cost of under-sampling large ones (which are very well sampled anyway). This feature becomes important especially when the category sizes differ significantly, which is the case in Facebook. Indeed, Fig. 5(a) shows that college sizes exhibit great heterogeneity. For a fair comparison, we only include the 5,331 colleges discovered by RW. (In fact, this filtering actually gives preference to RW. S-WRW crawls discovered many more colleges that we do not show in this figure.) They span more than two orders of magnitude and follow a heavily skewed distribution (not shown here).

Fig. 5(b) confirms that S-WRW successfully oversamples the small colleges. Indeed, the number of S-WRW samples

per college is almost constant (roughly around 100). In contrast, the number of RW samples follows closely the college size, which results in dramatic 100-fold differences between RW and S-WRW for smaller colleges.

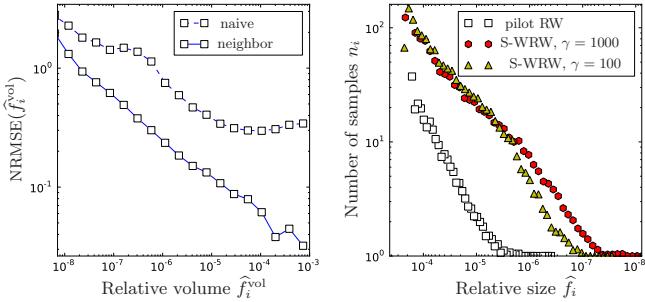
### 6.2.3 College size estimation

With more samples per college, we naturally expect a better estimation accuracy under S-WRW. We demonstrate it for three colleges of different sizes (in terms of the number of Facebook users): MIT (large), Caltech (medium), and Eindhoven University of Technology (small). Each boxplot in Fig. 5(c-e) is generated based on 25 independent college size estimates  $\hat{f}_i$  that come from walks of length  $n = 4K$  (left),  $n = 20K$  (middle), and  $n = 40K$  (right) samples each. For the three studied colleges, RW fails to produce reliable estimates in all cases except for MIT (largest college) under the two longest crawls. Similar results hold for the overwhelming majority of middle-sized and small colleges. The underlying reason is the very small number of samples collected by RW in these colleges, averaging at below 1 sample per walk. In contrast, the three S-WRW crawls contain typically 5-50 times more samples than RW (in agreement with Fig. 5(b)), and produce much more reliable estimates.

Finally, we aggregate the results over all colleges and compute the gain  $\alpha$  of S-WRW over RW. We calculate the error NRMSE( $\hat{f}_i$ ) by taking as our “ground truth”  $f_i$  the grand average of  $\hat{f}_i$  values over all samples collected via all full-length walks and crawl types. Fig. 5(f) presents NRMSE( $\hat{f}_i$ ) averaged over all 5,331 colleges discovered by RW, as a function of walk length  $n$ . As expected, for all crawl types the error decreases with  $n$ . However, there is a consistent large gap between RW and all three versions of S-WRW. RW needs 13-15 times more samples than S-WRW in order to achieve the same error.

### 6.2.4 The effect of the choice of $\gamma$

Recall that in all the S-WRW results described above, we used the resolution  $\gamma = 1000$ . In order to check how sensitive the results are to the choice of this parameter, we also tried a (shorter) S-WRW run with  $\gamma = 100$ , *i.e.*, ten times smaller. In Fig. 6(b), we see that the number of samples collected in the smallest colleges is smaller under  $\gamma = 100$  than under  $\gamma = 1000$ . In fact, the two curves diverge for colleges about



**Figure 6: Facebook: Pilot RW and other walks of the same length  $n = 65K$ .** (a) The performance of the neighbor-based volume estimator Eq.(35) (plain line) and the naive one Eq.(32) (dashed line). As ‘ground-truth’ we used  $f_i^{\text{vol}}$  calculated for all  $4 \times 1M$  collected samples. (b) The effect of the choice of  $\gamma$ .

100 times smaller than the biggest college, *i.e.*, exactly at the maximal resolution  $\gamma=100$ .

In any case, both settings of  $\gamma$  perform orders of magnitude better than RW of the same length.

### 6.3 Summary

Only about 3.5% of 500M Facebook users are college members. There are more than 10K colleges and they greatly vary in size, ranging from 50 (or fewer) to 50K members (we aggregate students, alumni and staff). In this setting, state-of-the-art sampling methods such as RW are bound to perform poorly. Indeed, UIS, *i.e.*, an idealized version of RW, with as many as 1M samples will collect only one sample from size-500 college, on average. Even if we could magically sample directly only from colleges, we would typically collect fewer than 30 samples per size-500 college.

S-WRW solves these problems. We showed that S-WRW of the same length collects typically about 100 samples per size-500 college. As a result, S-WRW outperforms RW by  $\alpha = 13 - 15$  times or  $\alpha = 12 - 14$  times if we also consider the 6.5% overhead from the initial pilot RW. This huge gain can be decomposed into two factors, say  $\alpha = \alpha_1 \cdot \alpha_2$ , as we proposed in Eq.(21). Factor  $\alpha_1 \simeq 8$  can be attributed to a about 8 times higher fraction of college samples in S-WRW compared to RW. Factor  $\alpha_2 \simeq 1.5$  is due to over-sampling smaller networks, *i.e.*, by applying stratified sampling.

Another important observation is that S-WRW is robust to the way we resolve target edge weight conflicts in Sec. 4.2.5. The differences between the three S-WRW implementations are minor - it is the application of Eq.(27) that brings most of the benefit.

## 7. RELATED WORK

**Graph Sampling by Exploration.** Early crawling of P2P, OSN and WWW typically used graph traversals, mainly BFS [3,31–33,43] and its variants. However, incomplete BFS introduces bias towards high-degree nodes that is unknown and thus impossible to correct in general graphs [2,8,17,25,26]. Later studies followed a more principled approach based on random walks (RW) [4,29]. The Metropolis-Hastings RW (MHRW) [16,30] removes the bias during the walk; it has been used to sample P2P networks [35,40] and OSNs [17]. Alternatively, we can use RW, whose bias is known and can be corrected for [20,39], thus leading to a re-weighted

RW [17,35]. RW was also used to sample Web [21], P2P networks [18,35,40], OSNs [17,24,33,36], and other large graphs [27]. It was empirically shown in [17,35] that RW outperforms MHRW in measurement accuracy. Therefore, RW can be considered as the state-of-the-art.

Random walks have also been used to sample *dynamic graphs* [35,40,42], which are outside the scope of this paper.

**Fast Mixing Markov Chains.** The mixing time of a random walk determines the efficiency of the sampling. On the practical side, the mixing time of RW in many OSNs was found larger than commonly believed [33]. Multiple dependent random walks [37] have been used to sample disconnected and loosely connected graphs. Random walks with jumps have been used to sample large graphs in [5,38] and in [27]. All the above methods treat all nodes with equal importance, which is orthogonal to our technique.

On the theoretical side, in [10], the authors propose a method to set edge weights that achieve the fastest mixing WRW for a given target stationary distribution. This technique, although related, is not applicable in our context. First, [10] requires the knowledge of the graph, which makes it inapplicable to  $G$ , yet possibly feasible in  $G^C$  (after estimating some limited information about  $G^C$  as in Sec. 4.2.1). In the latter case, however, even given a perfect knowledge of  $G^C$ , [10] often assigns weight 0 to some self-loops, which likely makes the underlying graph  $G$  disconnected. Finally, and most importantly, [10] takes a target stationary distribution as input. By taking  $w^{\text{WIS}}$ , we will face exactly the same problems of potentially poor convergence (Sec. 4.2.3) and “black holes” (Sec. 4.2.4) as we addressed by S-WRW.

**Stratified Sampling.** Our approach builds on *stratified sampling* [34], a widely used technique in statistics; see [12, 28] for a good introduction.

A related work in a different networking problem is [14], where threshold sampling is used to vary sampling probabilities of network traffic flows and estimate their volume.

**Weighted Random Walks for Sampling.** Random walks on graphs with weighted edges, or equivalently reversible Markov chains [4,29], are well studied and heavily used in Monte Carlo Markov Chain simulations [16] to sample a state space with a specified probability distribution. However, to the best of our knowledge, WRWs have not been designed explicitly for measurements of real online systems. In the context of sampling OSNs, the closest works are [5,38]. Technically speaking, they use WRW. But they set as their only objective the minimization of the mixing time, which makes them orthogonal and complementary to our approach, as we discussed above.

Very recent applications of weighted random walks in online social networks include [6,7]. [7] uses WRW in the context of link prediction. The authors employ supervised learning techniques to set the edge weights, with the goal of increasing the probability of visiting nodes that are more likely to receive new links. [6] introduces WRW-based methods to generate samples of nodes that are internally well-connected but also approximately uniform over the population. In both these papers, WRW is used to predict/extract something from a known graph. In contrast, we use WRW to estimate features of an unknown graph.

In the context of World Wide Web crawling, *focused crawling* techniques [11,13] have been introduced to follow web pages of specified interest and to avoid the irrelevant pages.

This is achieved by performing a BFS type of sample, except that instead of fifo queue they use a priority queue weighted by the page relevancy. In our context, such an approach suffers from the same problems as regular BFS: (i) collected samples strongly depend on the starting point, and (ii) we are not able to unbias the sample.

## 8. CONCLUSION

We introduced Stratified Weighted Random Walk (S-WRW) - an efficient way to sample large, static, undirected graphs via crawling and using minimal information. S-WRW performs a weighted random walk on the graph with weights determined by the estimation problem. We apply our approach to measure the Facebook social graph, and we show that S-WRW greatly outperforms the state-of-art sampling technique, namely the simple re-weighted random walk.

There are several directions for future work. First, S-WRW is currently an intuitive and efficient heuristic; in future work, we plan to investigate the optimal solution to problems identified in this paper and compare against or improve S-WRW. Second, it may be possible to combine these ideas with existing orthogonal techniques, some of which have been reviewed in Related Work, to further improve performance. Finally, we are interested in extending our techniques to dynamic graphs and non-stratified data.

## 9. REFERENCES

- [1] Weighted Random Walks of the Facebook social graph: <http://odysseas.calit2.uci.edu/research/>, 2011.
- [2] D. Achlioptas, A. Clauset, D. Kempe, and C. On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. *Journal of the ACM*, 2009.
- [3] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *WWW*, pages 835–844, 2007.
- [4] D. Aldous and J. A. Fill. *Reversible Markov Chains and Random Walks on Graphs*. In preparation.
- [5] K. Avrachenkov, B. Ribeiro, and D. Towsley. Improving Random Walk Estimation Accuracy with Uniform Restarts. In *17th Workshop on Algorithms and Models for the Web Graph*, 2010.
- [6] L. Backstrom and J. Kleinberg. Network Bucket Testing. In *WWW*, 2011.
- [7] L. Backstrom and J. Leskovec. Supervised Random Walks: Predicting and Recommending Links in Social Networks. In *ACM International Conference on Web Search and Data Mining (WSDM)*, 2011.
- [8] L. Becchetti, C. Castillo, D. Donato, and A. Fazzone. A comparison of sampling techniques for web graph characterization. In *LinkKDD*, 2006.
- [9] H. R. Bernard, T. Hallett, A. Iovita, E. C. Johnsen, R. Lyerla, C. McCarty, M. Mahy, M. J. Salganik, T. Saliuk, O. Scutelnicu, G. a. Shelley, P. Sirinirund, S. Weir, and D. F. Stroup. Counting hard-to-count populations: the network scale-up method for public health. *Sexually Transmitted Infections*, 86(Suppl 2):ii11–ii15, Nov. 2010.
- [10] S. Boyd, P. Diaconis, and L. Xiao. Fastest mixing Markov chain on a graph. *SIAM review*, 46(4):667–689, 2004.
- [11] S. Chakrabarti. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks*, 31(11-16):1623–1640, May 1999.
- [12] W. G. Cochran. *Sampling Techniques*, volume 20 of *McGraw-Hill Series in Probability and Statistics*. Wiley, 1977.
- [13] M. Diligenti, F. Coetze, S. Lawrence, C. Giles, and M. Gori. Focused crawling using context graphs. In *Proceedings of the 26th International Conference on Very Large Data Bases*, pages 527–534, 2000.
- [14] N. Duffield, C. Lund, and M. Thorup. Learn more, sample less: control of volume and variance in network measurement. *IEEE Transactions on Information Theory*, 51(5):1756–1775, May 2005.
- [15] J. Gentle. *Random number generation and Monte Carlo methods*. Springer Verlag, 2003.
- [16] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, 1996.
- [17] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In *INFOCOM*, 2010.
- [18] C. Gkantsidis, M. Mihail, and A. Saberi. Random walks in peer-to-peer networks. In *INFOCOM*, 2004.
- [19] M. Hansen and W. Hurwitz. On the Theory of Sampling from Finite Populations. *Annals of Mathematical Statistics*, 14(3), 1943.
- [20] D. D. Heckathorn. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems*, 44:174–199, 1997.
- [21] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near-uniform URL sampling. In *WWW*, 2000.
- [22] M. H. Kalos and P. A. Whitlock. *Monte carlo methods. Volume I: Basics*. Wiley, 1986.
- [23] E. D. Kolaczyk. *Statistical Analysis of Network Data*, volume 69 of *Springer Series in Statistics*. Springer New York, 2009.
- [24] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about Twitter. In *WOSN*, 2008.
- [25] M. Kurant, A. Markopoulou, and P. Thiran. On the bias of BFS (Breadth First Search). In *ITC*, also in *arXiv:1004.1729*, 2010.
- [26] S. H. Lee, P.-J. Kim, and H. Jeong. Statistical properties of Sampled Networks. *Phys. Rev. E*, 73:16102, 2006.
- [27] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *KDD*, pages 631–636, 2006.
- [28] S. Lohr. *Sampling: design and analysis*. Brooks/Cole, second edition, 2009.
- [29] L. Lovász. Random walks on graphs: A survey. *Combinatorics, Paul Erdos is Eighty*, 2(1):1–46, 1993.
- [30] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculation by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.
- [31] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the Flickr social network. In *WOSN*, 2008.
- [32] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel,

and B. Bhattacharjee. Measurement and analysis of online social networks. In *IMC*, pages 29–42, 2007.

[33] A. Mohaisen, A. Yun, and Y. Kim. Measuring the mixing time of social graphs. *IMC*, 2010.

[34] J. Neyman. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4):558, 1934.

[35] A. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach. Respondent-driven sampling for characterizing unstructured overlays. In *Infocom Mini-conference*, pages 2701–2705, 2009.

[36] A. H. Rasti, M. Torkjazi, R. Rejaie, and D. Stutzbach. Evaluating Sampling Techniques for Large Dynamic Graphs. In *Technical Report*, volume 1, 2008.

[37] B. Ribeiro and D. Towsley. Estimating and sampling graphs with multidimensional random walks. In *IMC*, volume 011, 2010.

[38] B. Ribeiro, P. Wang, and D. Towsley. On Estimating Degree Distributions of Directed Graphs through Sampling. *UMass Technical Report*, 2010.

[39] M. Salganik and D. D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34(1):193–240, 2004.

[40] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. In *IMC*, 2006.

[41] E. Volz and D. D. Heckathorn. Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, 24(1):79–97, 2008.

[42] W. Willinger, R. Rejaie, M. Torkjazi, M. Valafar, and M. Maggioni. OSN Research: Time to Face the Real Challenges. In *HotMetrics*, 2009.

[43] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. In *EuroSys*, 2009.

## Appendix A: Achieving Arbitrary Node Weights

Achieving arbitrary node weights by setting the edge weights in a graph  $G = (V, E)$  is sometimes impossible. For example, for a graph that is a path consisting of two nodes  $(v_1 - v_2)$ , it is impossible to achieve  $w(v_1) \neq w(v_2)$ . However, it is always possible to do so, if there are self loops in each node.

**OBSERVATION 1.** For any undirected graph  $G = (V, E)$  with a self-loop  $\{v, v\}$  at every node  $v \in V$ , we can achieve an arbitrary distribution of node weights  $w(v) > 0$ ,  $v \in V$ , by appropriate choice of edge weights  $w(u, v) > 0$ ,  $\{u, v\} \in E$ .

**PROOF.** Denote by  $w_{\min}$  the smallest of all target node weights  $w(v)$ . Set  $w(u, v) = w_{\min}/N$  for all non self-loop edges (i.e., where  $u \neq v$ ). Now, for every self-loop  $\{v, v\} \in E$  set

$$w(v, v) = \frac{1}{2} \left( w(v) - \frac{w_{\min}}{N} \cdot (\deg(v) - 2) \right).$$

It is easy to check that, because there are exactly  $\deg(v) - 2$  non self-loop edges incident on  $v$ , every node  $v \in V$  will achieve the target weight  $w(v)$ . Moreover, the definition of  $w_{\min}$  guarantees that  $w(v, v) > 0$  for every  $v \in V$ .  $\square$

## Appendix B: Estimating Category Volumes

In this section, we derive efficient estimators of the volume ratio  $\hat{f}_C^{\text{vol}} = \frac{\text{vol}(C)}{\text{vol}(V)}$ . Recall that  $S \subset V$  denotes an independent sample of nodes in  $G$ , with replacement.

### Node sampling

If  $S$  is a uniform sample UIS, then we can write

$$\hat{f}_C^{\text{vol}} = \frac{\sum_{v \in S} \deg(v) \cdot 1_{\{v \in C\}}}{\sum_{v \in S} \deg(v)}, \quad (30)$$

which is a straightforward application of the classic ratio estimator [28].

In the more general case, when  $S$  is selected using WIS, then we have to correct for the linear bias towards nodes of higher weights  $w()$ , as follows:

$$\hat{f}_C^{\text{vol}} = \frac{\sum_{v \in S} \deg(v) \cdot 1_{\{v \in C\}} / w(v)}{\sum_{v \in S} \deg(v) / w(v)}. \quad (31)$$

In particular, if  $w(v) \sim \deg(v)$ , then

$$\hat{f}_C^{\text{vol}} = \frac{1}{n} \cdot \sum_{v \in S} 1_{\{v \in C\}}. \quad (32)$$

### Star sampling

Another approach is to focus on the set of all neighbors  $\mathcal{N}(S)$  of sampled nodes (with repetitions) rather than on  $S$  itself, i.e., to use ‘star sampling’ [23]. The probability that a node  $v$  is a neighbor of a node sampled from  $V$  by UIS is

$$\sum_{u \in V} \frac{1}{N} \cdot 1_{\{v \in \mathcal{N}(u)\}} = \frac{\deg(v)}{N}.$$

Consequently, the nodes in  $\mathcal{N}(S)$  are asymptotically equivalent to nodes drawn with probabilities linearly proportional to node degrees. By applying Eq.(32) to  $\mathcal{N}(S)$ , we obtain<sup>9</sup>

$$\hat{f}_C^{\text{vol}} = \frac{1}{\text{vol}(S)} \sum_{u \in S} \sum_{v \in \mathcal{N}(u)} 1_{\{v \in C\}}, \quad (33)$$

where we used  $|\mathcal{N}(S)| = \sum_{u \in S} \deg(u) = \text{vol}(S)$ .

In the more general case, when  $S$  is selected using WIS, then we correct for the linear bias towards nodes of higher weights  $w()$ , as follows:

$$\hat{f}_C^{\text{vol}} = \frac{1}{\sum_{u \in S} \frac{\deg(u)}{w(u)}} \sum_{u \in S} \left( \frac{1}{w(u)} \sum_{v \in \mathcal{N}(u)} 1_{\{v \in C\}} \right). \quad (34)$$

In particular, if  $w(v) \sim \deg(v)$ , then

$$\hat{f}_C^{\text{vol}} = \frac{1}{n} \sum_{u \in S} \left( \frac{1}{\deg(u)} \sum_{v \in \mathcal{N}(u)} 1_{\{v \in C\}} \right). \quad (35)$$

Note that for every sampled node  $v \in S$ , the formulas Eq.(33-35) exploit all the  $\deg(v)$  neighbors of  $v$ , whereas Eq.(30-32) rely on one node per sample only. Not surprisingly, Eq.(33-35) performed much better in all our simulations and implementations.

<sup>9</sup>As a side note, observe that formula Eq.(33) generalizes the ‘scale-up method’ [9] used in social sciences to estimate the size (here  $|C|$ ) of hidden populations (e.g., of drug addicts). Indeed, if we assume that the average node degree in  $V$  is the same as in  $C$ , then  $\text{vol}(C)/\text{vol}(V) = |C|/N$ , which reduces Eq.(32) to the core formula of the scale-up method.

## Appendix C: Relative sizes of node categories

Consider a scenario with only two node categories, i.e.,  $\mathcal{C} = \{C_1, C_2\}$ . Denote  $f_1 = |C_1|/N$  and  $f_2 = |C_2|/N$ . The goal is to estimate  $f_1$  and  $f_2$  based on the collected sample  $S$ .

### UIS - Uniform independence sampling.

Under UIS, the number  $X_1$  of times we select a node from  $C_1$  among  $n$  attempts follows the Binomial distribution  $X_1 = \text{Binom}(f_1, n)$ . Therefore, we can estimate  $f_1$  as

$$\hat{f}_1^{\text{UIS}} = \frac{X_1}{n} \quad \text{with} \quad \mathbb{V}(\hat{f}_1^{\text{UIS}}) = \frac{f_1 f_2}{n}. \quad (36)$$

### WIS - Weighted independence sampling.

In contrast, under WIS, at every iteration the probability  $\pi(v)$  of selecting a node  $v$  is:

$$\pi(v) = \begin{cases} \pi_1 = \frac{1}{N} \cdot \frac{w_1}{w_1 f_1 + w_2 f_2} & \text{if } v \in C_1, \text{ and} \\ \pi_2 = \frac{1}{N} \cdot \frac{w_2}{w_1 f_1 + w_2 f_2} & \text{if } v \in C_2, \end{cases}$$

where  $w_1$  and  $w_2$  are the weights  $w(v)$  of nodes in  $C_1$  and  $C_2$ , respectively.

By applying the Hansen-Hurwitz estimator (separately for nominator and denominator), we obtain

$$\begin{aligned} \hat{f}_1^{\text{WIS}} &= \frac{|\hat{C}_1|}{\hat{N}} = \frac{\sum_{v \in S} 1_{v \in C_1} / \pi(v)}{\sum_{v \in S} 1 / \pi(v)} \\ &= \frac{X_1 / \pi_1}{X_1 / \pi_1 + (n - X_1) / \pi_2} \\ &= \frac{X_1 \cdot \pi_2}{X_1(\pi_2 - \pi_1) + n \cdot \pi_1} \\ &= \frac{X_1 \cdot w_2}{X_1(w_2 - w_1) + n \cdot w_1}, \end{aligned} \quad (37)$$

where  $X_1$  is the number of samples taken from  $C_1$ . Note, that to calculate  $\hat{f}_1^{\text{WIS}}$  we only need values  $w_1$  and  $w_2$ , which are set by us and thus known.

Computing the variance of  $\hat{f}_1^{\text{WIS}}$  is a bit more challenging. We use the second-order Taylor expansions (the 'Delta method') to approximate it as follows:

$$\begin{aligned} \frac{\partial \hat{f}_1^{\text{WIS}}}{\partial X_1} &= \frac{n w_1 w_2}{((w_2 - w_1) X_1 + n w_1)^2}, \quad \text{and} \\ \mathbb{V}(\hat{f}_1^{\text{WIS}}) &\cong \left( \frac{\partial \hat{f}_1^{\text{WIS}}}{\partial X_1} (\mathbb{E}(X_1)) \right)^2 \mathbb{V}(X_1) \\ &= (\dots) = \frac{f_1 f_2}{n w_1 w_2} \cdot (f_1 w_1 + f_2 w_2)^2. \end{aligned} \quad (38)$$

In the above derivation, we used the fact that  $\mathbb{E}(X_1) = n N f_1 \pi_1$  and  $\mathbb{V}(X_1) = n N^2 f_1 \pi_1 f_2 \pi_2$ . This comes from the fact that  $X_1$  actually follows the binomial distribution  $X_1 = \text{Binom}(N f_1 \pi_1, n)$ .

For  $w_1 = w_2$ , we are back in the UIS case. But this is not necessarily the optimal choice of weights. Indeed, a quick application of Lagrange multipliers reveals that  $\mathbb{V}(\hat{f}_1^{\text{WIS}})$  is minimized when

$$w_1 f_1 = f_2 w_2. \quad (39)$$

Moreover, analogous analysis shows that Eq.(39) minimizes  $\mathbb{V}(\hat{f}_2^{\text{WIS}})$  as well. In other words, the estimators of both  $f_1$  and  $f_2$  have the lowest variance if the total weighted mass

of  $C_1$  is equal to that of  $C_2$ . This implies, in expectation, equal allocation of samples between  $C_1$  and  $C_2$ , i.e.,

$$n_i^{\text{WIS}} = \frac{n}{|\mathcal{C}|}.$$

Finally, we can use Eq.(36), Eq.(38) and Eq.(39) to calculate the gain  $\alpha$  of WIS over UIS

$$\alpha = \frac{1}{4 f_1 f_2} \quad (\geq 1). \quad (40)$$

Note that we always have  $\alpha \geq 1$ , and  $\alpha$  grows quickly with growing difference between  $f_1$  and  $f_2$ .

## Appendix D: Optimal WRW weights in Fig. 3(a)

Every time WRW visits the white node/category in Fig. 3(a), the next node is chosen uniformly from red and green categories. We stay in this selected category for  $k$  rounds, where  $k$  is a geometric random variable with parameter  $p = w_2/(w_1+w_2) \in [0, 1]$ . Next, we come back to the white category, and reiterate the process. So the number  $n_{\text{red}}$  of times the red category is sampled is

$$n_{\text{red}} = \sum_1^{\text{Binom}(0.5, n_{\text{wh}})} \text{Geom}(p),$$

where  $n_{\text{wh}}$  is the number of visits to the white category. Because the random variables generated by  $\text{Binom}(0.5, n_{\text{wh}})$  and  $\text{Geom}(p)$  are independent, we can write

$$\begin{aligned} \mathbb{E}[n_{\text{red}}] &= \mathbb{E}[\text{Binom}(0.5, n_{\text{wh}})] \cdot \mathbb{E}[\text{Geom}(p)] = 0.5 n_{\text{wh}} / p \\ \mathbb{V}[n_{\text{red}}] &= \mathbb{E}[\text{Binom}()] \mathbb{V}[\text{Geom}()] + \mathbb{E}^2[\text{Geom}()] \mathbb{V}[\text{Binom}()] \\ &= \frac{n_{\text{wh}}}{4p^2} (3 - 2p). \end{aligned}$$

A possible unbiased estimator of the relative size  $f_{\text{red}}$  of red category (among relevant categories) is

$$\hat{f}_{\text{red}} = \frac{n_{\text{red}}}{n_{\text{wh}} / p},$$

for which we get

$$\begin{aligned} \mathbb{E}[\hat{f}_{\text{red}}] &= \frac{\mathbb{E}[n_{\text{red}}]}{n_{\text{wh}} / p} = \frac{1}{2} \quad (\text{unbiased}) \\ \mathbb{V}[\hat{f}_{\text{red}}] &= \frac{\mathbb{V}[n_{\text{red}}]}{(n_{\text{wh}} / p)^2} = \frac{3 - 2p}{4n_{\text{wh}}}. \end{aligned}$$

This variance is expressed as a function of  $n_{\text{wh}}$ , and not of the total sample length  $n$ . However, note that  $n_{\text{wh}}$  drops with decreasing  $p$ . Consequently, the variance  $\mathbb{V}[\hat{f}_{\text{red}}]$  (expressed as a function of  $n_{\text{wh}}$  or of  $n$ ) is minimized for  $p = 1$ , i.e., for  $w_1 = 0$  and  $w_2 > 0$  (and  $n_{\text{wh}} = n/2$ ).